Speech technology and speech recognition for Luxembourgish

...

Peter Gilles
@PeterGilles
Institute for Luxembourgish linguistics and literatures | Department of Humanities
University of Luxembourg

Overview

1. Luxembourg - language situation

2. Research on speech technology at the University of Luxembourg

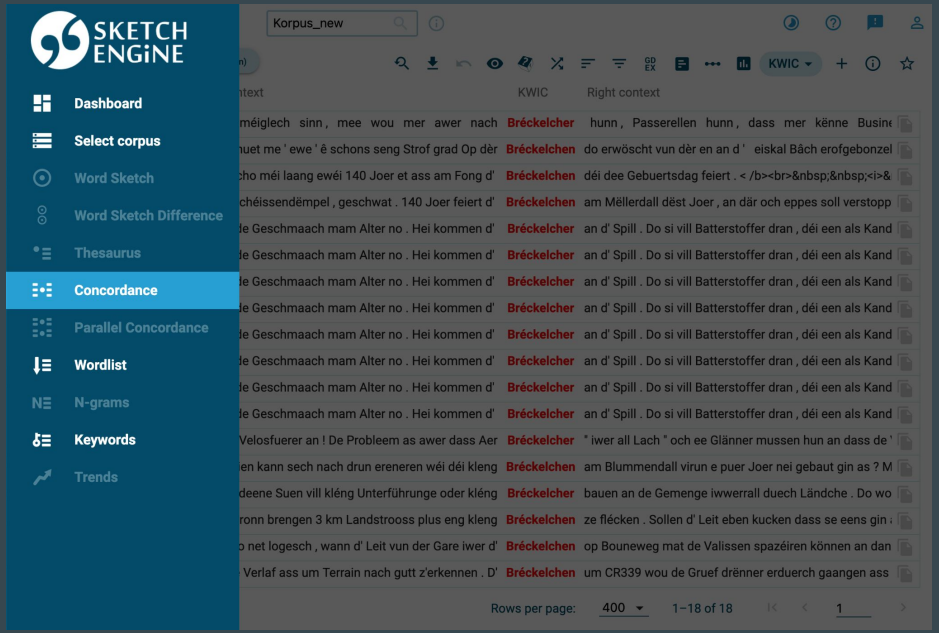
3. A future project on speech recognition

Luxembourg - language situation

- highly multilingual country
  - Luxembourgish (Lëtzebuergesch) as national language
    - mainly spoken language, for all informal and formal purposes
    - increased use as written language
    - ongoing codification (orthography, lexicon)
  - French
    - prestigious written language
    - language of the workplace
  - German
    - language of alphabetisation
    - receptive language of media consumption (newspapers, German TV)
- 620,000 inhabitants; 180,000 daily commuters; 300,000 - 400,000 speak Luxembourgish as 1st or 2nd language
- highly variable language
- not an official language of the EU (yet)

Speech technology @ University of Luxembourg

- since 2006 continuous development of tools and data
- text corpus
  - various sources: parliamentary debates/speeches, RTL news data, literature, social media ...
  - 100 Mio tokens
  - spelling normalisation, lemmatisation, PoS-tagged
  - indexed as NoSketch Engine



spellux -Automatic text normalization for Luxembourgish

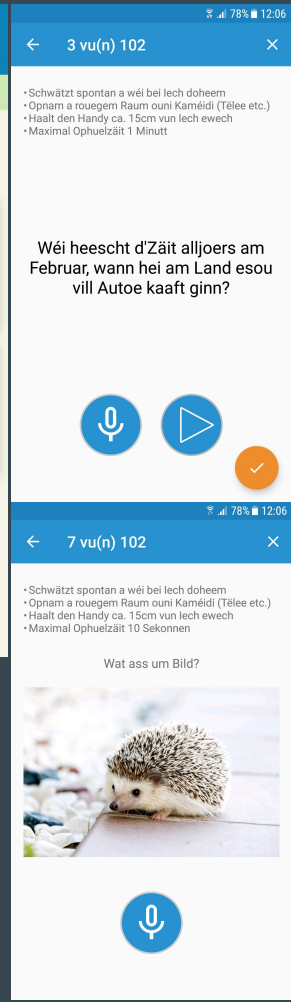
- preprocessing of text data to automatically correct spelling mistakes

Input: Eche hun d'Wort heut den Muaren mussen leesen.
Output: Ech hunn d'Wuert haut de Muere musse liesen.

- needed to improve the text corpus and the language models
- developed by Christoph Purschke; https://github.com/questoph/spellux

Creation of audio corpus

- collection with smartphone application ‘Schnëssen’
- self-recording of translation and picture naming tasks
- 300,000 recordings by 400 to 2500 speakers
- structured database of words and phonetic segments
- foreseen also as training for an acoustical model



Grapheme-to-phoneme conversion (G2P)

- developed to create a pronunciation dictionary of Luxembourgish
- some 20,000 manually phonetically transcribed words
- trained with ‘sequitur’; https://github.com/sequitur-g2p
- additional training for syllable detection; https://github.com/PeterGilles/Syllabifier-for-Luxembourgish

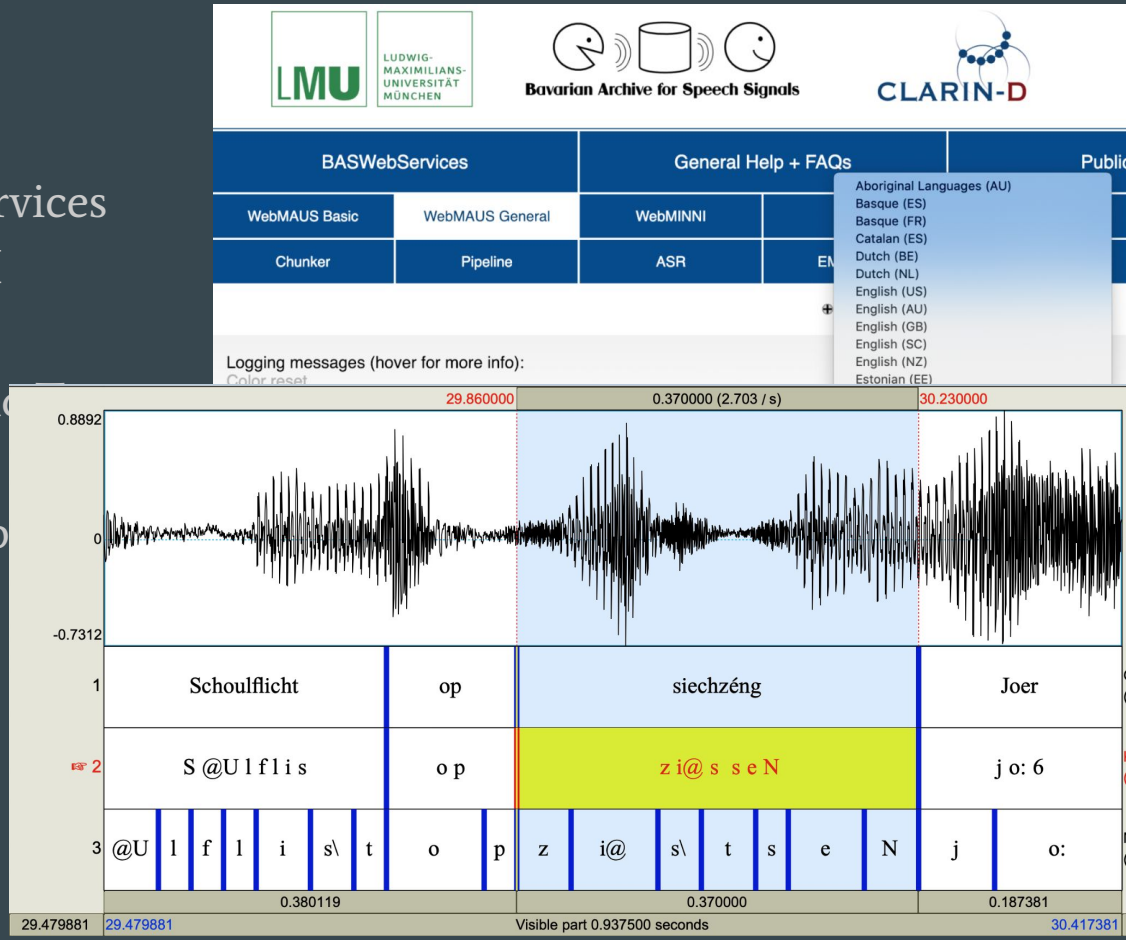
Text input
Um Sonndeg war den 100. Tour de Flandres an der Belsch. De Weltmeeschter Peter Sagan wënnt déi 100. Tour de Flandres.

IPA output
[um zondeɕ vaːʀ dən eːnhonɐt tuːʀ də flɑndʀəs ɑn dɐ bælʃ də væltmeːʃtɐ peːtɐ zɑgaːn vənt dɜɪ eːnhonɐt tuːʀ də flɑndʀəs]

IPA output (syllabified)
[um zon.deɕ vaːʀ dən eːn.ho.nɐt tuːʀ də flɑn.dʀəs ɑn dɐ bælʃ də vælt.meːʃ.tɐ peː.tɐ zɑ.gaːn vənt dɜɪ eːn.ho.nɐt tuːʀ də flɑn.dʀəs]

Forced alignment

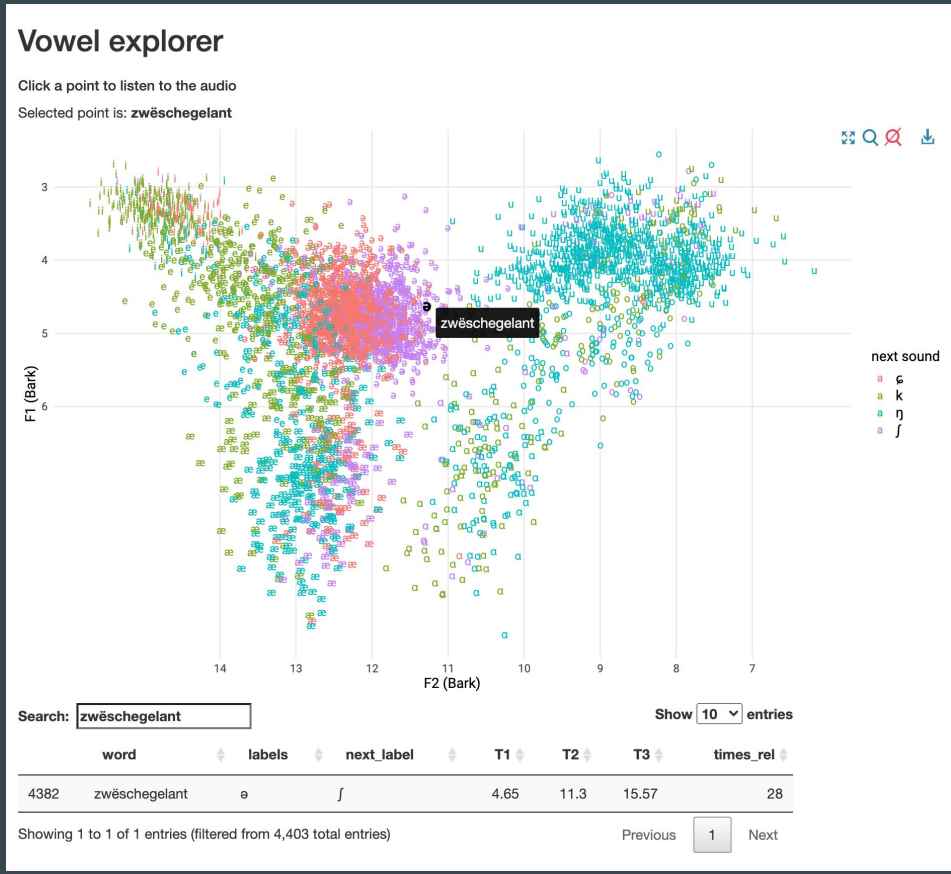
- collaboration with BASWebServices
- available as web service or API
- Input
  Audio + matching orthographi
- Output
  temporal alignment informatio
  words and phonetic segments



Forced alignment

- large-scale database of phonetic segments
- research e.g. in the structure and variability of the vowel system

- https://github.com/PeterGilles/vowel_explorer



Speech recognition (project in preparation)

- Partners
  - Luxembourgish Parliament
  - Center for Luxembourgish Language
  - University of Luxembourg
- development system for the automatic transcription of speeches and debates in the parliament
- training data
  - tons of audio/video recordings and written transcripts used to create the acoustical model
  - pronunciation lexicon provided by the UL
  - language model based on text corpora from the UL, the parliament
- based on an KALDI recipe

Challenges

- highly variable spoken language
  - generating enough training data
- receiving industry support for case studies
- reaching critical mass
  - finding competent & interested researchers